Maximum A Posteriori Estimation

Nipun Batra

August 21, 2023

IIT Gandhinagar

Agenda

Revision

Coin Toss Problem

MAP for Logistic Regression

Revision

Bayes Rule

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)}$$

- $P(\theta|D)$ is called the posterior
- $P(D|\theta)$ is called the likelihood
- $P(\theta)$ is called the prior
- $P(D)$ is called the evidence

Maximum Likelihood Estimation

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)} = \frac{P(D|\theta) \cdot P(\theta)}{\int_\theta P(D|\theta) \cdot P(\theta) d\theta}$$

Given a dataset D, find the parameters θ that maximize the likelihood of the data.

$$\theta_{\text{MLE}} = \arg\max_\theta P(D|\theta)$$

For example, given a linear regression problem setup, we set the likelihood as normal distribution and find the parameters θ that maximize the likelihood of the data.

Maximum A Posteriori Estimation

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)} = \frac{P(D|\theta) \cdot P(\theta)}{\int_\theta P(D|\theta) \cdot P(\theta) d\theta}$$

Given a dataset D, find the parameters θ that maximize the posterior of θ considering both the likelihood and the prior.

$$\theta_{\text{MAP}} = \arg\max_\theta P(\theta|D) = \arg\max_\theta P(D|\theta) \cdot P(\theta)$$

Maximum A Posteriori Estimation

- **MLE**: Given N observations, obtain best θ estimate (or θ_{MLE})

Maximum A Posteriori Estimation

- **MLE**: Given N observations, obtain best θ estimate (or θ_{MLE})
- What if we have prior knowledge about θ?

Maximum A Posteriori Estimation

- **MLE**: Given N observations, obtain best θ estimate (or θ_{MLE})
- What if we have prior knowledge about θ?
- **MAP**: Given N observations and prior knowledge, obtain best θ estimate (or θ_{MAP})

Maximum A Posteriori Estimation

- **MLE**: Given N observations, obtain best θ estimate (or θ_{MLE})
- What if we have prior knowledge about θ?
- **MAP**: Given N observations and prior knowledge, obtain best θ estimate (or θ_{MAP})
- When do we need prior knowledge?

Maximum A Posteriori Estimation

- **MLE**: Given N observations, obtain best θ estimate (or θ_{MLE})
- What if we have prior knowledge about θ?
- **MAP**: Given N observations and prior knowledge, obtain best θ estimate (or θ_{MAP})
- When do we need prior knowledge?
  - When the dataset is not a good representation of the true distribution.
  - Can be a data quality and/or quantity issue.

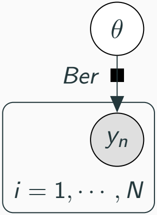
Coin Toss Problem

Coin Toss Problem

- Consider a sequence of independent N coin toss outcomes, $D = \{y_1, ..., y_N\}$ where each observation y_i is a binary random variable (Heads: 1, Tails: 0).

Coin Toss Problem

- Consider a sequence of independent N coin toss outcomes, $D = \{y_1, ..., y_N\}$ where each observation y_i is a binary random variable (Heads: 1, Tails: 0).
- Assuming $y_i \sim \text{Bernoulli}\,(\theta)$, $P(y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$



Coin Toss Problem

- For the sequence: $P(D|\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i}$

Coin Toss Problem

- For the sequence: $P(D|\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i}$
- **Recall**: $P(D|\theta) \longrightarrow$ Likelihood or $\mathcal{L}(\theta)$

Coin Toss Problem

- For the sequence: $P(D|\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i}$
- **Recall**: $P(D|\theta) \longrightarrow$ Likelihood or $\mathcal{L}(\theta)$
- Log-Likelihood or $\mathcal{LL}(\theta) = \sum_{i=1}^{N} y_i \log\theta + (1-y_i)\log(1-\theta)$

Coin Toss Problem

- For the sequence: $P(D|\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i}$
- **Recall**: $P(D|\theta) \longrightarrow$ Likelihood or $\mathcal{L}(\theta)$
- Log-Likelihood or $\mathcal{LL}(\theta) = \sum_{i=1}^{N} y_i \log\theta + (1-y_i)\log(1-\theta)$
- **Recall**: $\theta_{\text{MLE}} = \arg\max_{\theta} P(D|\theta)$

$$\therefore \frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0 \implies \theta_{MLE} = \frac{\sum_{i=1}^{N} y_i}{N}$$

Coin Toss Problem

- For the sequence: $P(D|\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i}$
- **Recall**: $P(D|\theta) \longrightarrow$ Likelihood or $\mathcal{L}(\theta)$
- Log-Likelihood or $\mathcal{LL}(\theta) = \sum_{i=1}^{N} y_i \log\theta + (1-y_i)\log(1-\theta)$
- **Recall**: $\theta_{\text{MLE}} = \arg\max_\theta P(D|\theta)$

$$\therefore \frac{\partial \mathcal{L}(\theta)}{\partial\theta} = 0 \implies \theta_{MLE} = \frac{\sum_{i=1}^{N} y_i}{N}$$

- Rewrite, $\theta_{MLE} = \frac{n_H}{n_H+n_T}$

Coin Toss Problem

- For the sequence: $P(D|\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i}$
- **Recall**: $P(D|\theta) \longrightarrow$ Likelihood or $\mathcal{L}(\theta)$
- Log-Likelihood or $\mathcal{LL}(\theta) = \sum_{i=1}^{N} y_i \log\theta + (1-y_i)\log(1-\theta)$
- **Recall**: $\theta_{\text{MLE}} = \arg\max_\theta P(D|\theta)$

$$\therefore \frac{\partial\mathcal{L}(\theta)}{\partial\theta} = 0 \implies \theta_{MLE} = \frac{\sum_{i=1}^{N} y_i}{N}$$

- Rewrite, $\theta_{MLE} = \frac{n_H}{n_H+n_T}$
- Suppose 10 tosses yield 9 heads and 1 tail. $\theta_{MLE} =$

Coin Toss Problem

- For the sequence: $P(D|\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i}$
- **Recall**: $P(D|\theta) \longrightarrow$ Likelihood or $\mathcal{L}(\theta)$
- Log-Likelihood or $\mathcal{LL}(\theta) = \sum_{i=1}^{N} y_i \log\theta + (1-y_i)\log(1-\theta)$
- **Recall**: $\theta_{\text{MLE}} = \arg\max_{\theta} P(D|\theta)$

$$\therefore \frac{\partial\mathcal{L}(\theta)}{\partial\theta} = 0 \implies \theta_{MLE} = \frac{\sum_{i=1}^{N} y_i}{N}$$

- Rewrite, $\theta_{MLE} = \frac{n_H}{n_H+n_T}$
- Suppose 10 tosses yield 9 heads and 1 tail. $\theta_{MLE} = 0.9$

Coin Toss Problem

- For the sequence: $P(D|\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i}$
- **Recall**: $P(D|\theta) \longrightarrow$ Likelihood or $\mathcal{L}(\theta)$
- Log-Likelihood or $\mathcal{LL}(\theta) = \sum_{i=1}^{N} y_i \log\theta + (1-y_i)\log(1-\theta)$
- **Recall**: $\theta_{\text{MLE}} = \arg\max_\theta P(D|\theta)$

$$\therefore \frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0 \implies \theta_{MLE} = \frac{\sum_{i=1}^{N} y_i}{N}$$

- Rewrite, $\theta_{MLE} = \frac{n_H}{n_H + n_T}$
- Suppose 10 tosses yield 9 heads and 1 tail. $\theta_{MLE} = 0.9$
- What if we have prior knowledge that the coin is fair?

Incorporating Prior Information

- We can incorporate prior information by assuming a prior distribution over θ.

Incorporating Prior Information

- We can incorporate prior information by assuming a prior distribution over θ.

$$\because P(\mathit{Head}) = \theta \in [0, 1]$$

Incorporating Prior Information

- We can incorporate prior information by assuming a prior distribution over θ.

$$\because P(Head) = \theta \in [0, 1]$$

- A resonable choice for prior is the Beta distribution.

Incorporating Prior Information

- We can incorporate prior information by assuming a prior distribution over θ.

$$\because P(Head) = \theta \in [0,1]$$

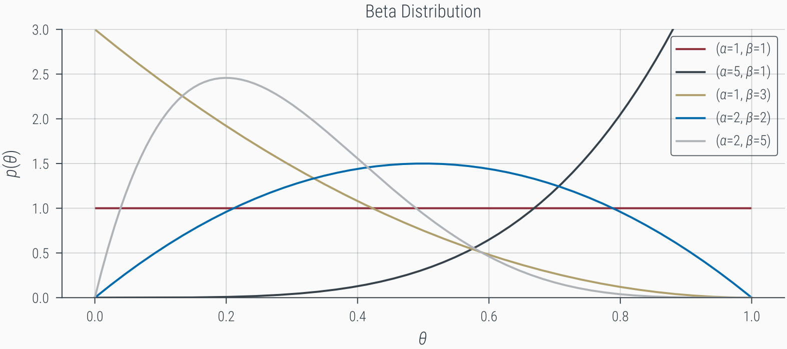
- A resonable choice for prior is the Beta distribution.

$$\implies P(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

where,

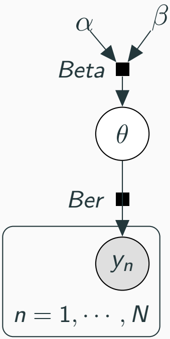
$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt \;\; \text{(Gamma Function)}$$

Beta Distribution



Notebook

Coin Toss Problem with Prior



Deriving θ_{MAP}

- **Recall**: $\theta_{\text{MAP}} = \arg\max_\theta P(\theta|D) = \arg\max_\theta P(D|\theta) \cdot P(\theta)$

Deriving θ_{MAP}

- **Recall**: $\theta_{\text{MAP}} = \arg\max_\theta P(\theta|D) = \arg\max_\theta P(D|\theta) \cdot P(\theta)$
- The log-posterior for this coin-toss problem is given as,

Deriving θ_{MAP}

- **Recall**: $\theta_{\text{MAP}} = \arg\max_\theta P(\theta|D) = \arg\max_\theta P(D|\theta) \cdot P(\theta)$
- The log-posterior for this coin-toss problem is given as,

$$\log P(\theta|D) = \sum_{i=1}^{N} \log P(y_i|\theta) + \log P(\theta)$$

Deriving θ_{MAP}

- **Recall**: $\theta_{\text{MAP}} = \arg\max_\theta P(\theta|D) = \arg\max_\theta P(D|\theta) \cdot P(\theta)$
- The log-posterior for this coin-toss problem is given as,

$$\log P(\theta|D) = \sum_{i=1}^{N} \log P(y_i|\theta) + \log P(\theta)$$

$$\log P(\theta|D) = \sum_{i=1}^{N} y_i \log\theta + (1-y_i)\log(1-\theta) + \\ (\alpha-1)\log\theta + (\beta-1)\log(1-\theta)$$

Deriving θ_{MAP}

$$\frac{\partial \log P(\theta|D)}{\partial \theta} = \frac{\sum_{i=1}^{N} y_i}{\theta} - \frac{\sum_{i=1}^{N}(1-y_i)}{1-\theta} + \frac{\alpha-1}{\theta} - \frac{\beta-1}{1-\theta} = 0$$

$$\implies (1-\theta)\sum_{i=1}^{N} y_i + \theta\sum_{i=1}^{N}(1-y_i) + (1-\theta)(\alpha-1) - \theta(\beta-1) = 0$$

$$\implies \sum_{i=1}^{N} y_i - \theta\sum_{i=1}^{N} y_i - N\theta + \theta\sum_{i=1}^{N} y_i + \alpha - 1 - \theta\alpha + \theta - \theta\beta + \theta = 0$$

Deriving θ_{MAP}

$$\implies \sum_{i=1}^{N} y_i + \alpha - 1 - \theta(N + \alpha + \beta - 2) = 0$$

$$\implies \theta_{MAP} = \frac{\sum_{i=1}^{N} y_i + \alpha - 1}{N + \alpha + \beta - 2}$$

Deriving θ_{MAP} (Coin toss context)

- Total number of tosses $= N$

Deriving θ_{MAP} (Coin toss context)

- Total number of tosses $= N$
- Number of heads $(y = 1) = n_H$

Deriving θ_{MAP} (Coin toss context)

- Total number of tosses $= N$
- Number of heads $(y = 1) = n_H$
- Number of tails $(y = 0)$= n_T

Deriving θ_{MAP} (Coin toss context)

- Total number of tosses $= N$
- Number of heads $(y = 1) = n_H$
- Number of tails $(y = 0)= n_T$
- Pseudo heads $= \alpha$

Deriving θ_{MAP} (Coin toss context)

- Total number of tosses $= N$
- Number of heads $(y = 1) = n_H$
- Number of tails $(y = 0) = n_T$
- Pseudo heads $= \alpha$
- Pseudo tails $= \beta$

Deriving θ_{MAP} (Coin toss context)

- Total number of tosses $= N$
- Number of heads $(y = 1) = n_H$
- Number of tails $(y = 0) = n_T$
- Pseudo heads $= \alpha$
- Pseudo tails $= \beta$
- $\theta_{MAP} = \frac{n_H + \alpha - 1}{N + \alpha + \beta - 2}$

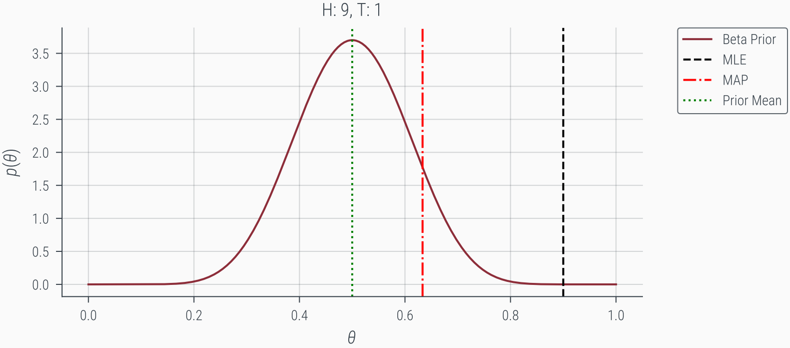
Deriving θ_{MAP} (Coin toss context)

- Total number of tosses $= N$
- Number of heads $(y = 1) = n_H$
- Number of tails $(y = 0) = n_T$
- Pseudo heads $= \alpha$
- Pseudo tails $= \beta$
- $\theta_{MAP} = \frac{n_H + \alpha - 1}{N + \alpha + \beta - 2}$
- Prior $=$ Beta(α, β)

Deriving θ_{MAP} (Coin toss context)

- Total number of tosses $= N$
- Number of heads $(y = 1) = n_H$
- Number of tails $(y = 0)= n_T$
- Pseudo heads $= \alpha$
- Pseudo tails $= \beta$
- $\theta_{MAP} = \frac{n_H+\alpha-1}{N+\alpha+\beta-2}$
- Prior $= \text{Beta}(\alpha, \beta)$
- Posterior $= \text{Beta}(n_H + \alpha, n_T + \beta)$

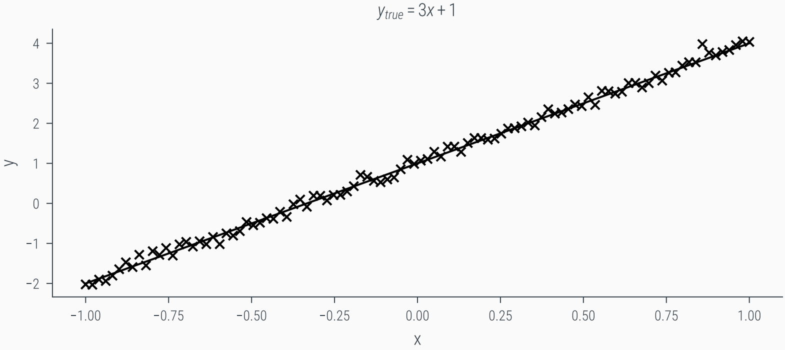
Coin Toss Problem with Prior



Notebook

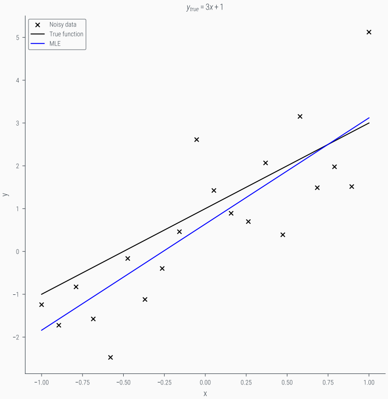
MLE for Linear Regression

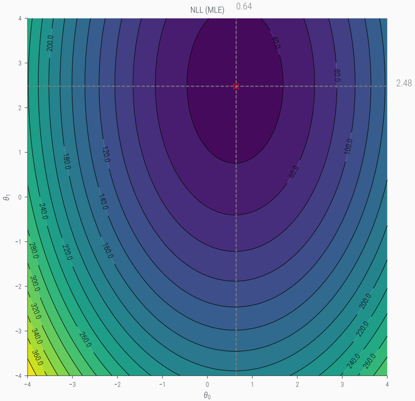
- Consider a dataset $D = \{(x_1, y_1)...(x_N, y_N)\}$ where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$.
- Suppose the data is generated from a linear model with additive Gaussian noise, i.e., $y_i = \theta^T x_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.



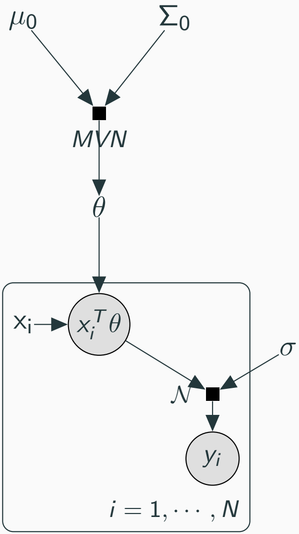
MLE for Linear Regression

- The likelihood is given by, $P(y_i|x_i,\theta) = \mathcal{N}(y_i|\theta^T x_i, \sigma^2)$
- **Recall**: The negative log-likelihood is given by, $\mathcal{NLL}(\theta) = \frac{1}{2\sigma^2}\left(y - X\theta\right)^T\left(y - X\theta\right)$
- **Recall**: The MLE is given by, $\theta_{MLE} = \arg\min_\theta \mathcal{NLL}(\theta) = \left(X^T X\right)^{-1} X^T y$





MAP for Linear Regression



MAP for Linear Regression

As per Bayes' rule,

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

MAP for Linear Regression

As per Bayes' rule,

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

- Log-Likelihood:
  $\mathcal{LL}(\theta) = \log P(D|\theta) = \log \prod_{i=1}^{N} \mathcal{N}(y_i|x_i^T\theta, \sigma^2)$
- Prior: $P(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Log-Prior: $\log P(\theta) = \log \mathcal{N}(\theta|\mu_0, \Sigma_0)$
- Log-Joint: $\log P(\theta|D) = \log P(D|\theta) + \log P(\theta)$

MAP for Linear Regression

Notebook

MAP for Linear Regression

$$\theta_{MAP} = \arg\min \log P(\theta|D) = \arg\min \mathcal{NLL}(\theta) + \log P(\theta)$$

Using zero-mean Gaussian prior

$P(\theta) = MVN(\mu_0, \Sigma_0)$

Using zero-mean Gaussian prior

$P(\theta) = MVN(\mu_0, \Sigma_0)$

Assume: $\mu_0 = \vec{0}$, $\Sigma_0 = \sigma_0^2 \mathsf{I}$

Using zero-mean Gaussian prior

$P(\theta) = MVN(\mu_0, \Sigma_0)$

Assume: $\mu_0 = \vec{0}$, $\Sigma_0 = \sigma_0^2 \mathrm{I}$

$$\theta_{MAP} = \arg\min \log P(\theta|D) = \arg\min \mathcal{NLL}(\theta) + \log P(\theta)$$

Using zero-mean Gaussian prior

$P(\theta) = MVN(\mu_0, \Sigma_0)$

Assume: $\mu_0 = \vec{0}$, $\Sigma_0 = \sigma_0^2 \mathrm{I}$

$$\theta_{MAP} = \arg\min \log P(\theta|D) = \arg\min \mathcal{NLL}(\theta) + \log P(\theta)$$

We get

$$\theta_{MAP} = \arg\min \frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta) + \frac{1}{\sigma_0^2}\theta^T\theta$$

Using zero-mean Gaussian prior

$P(\theta) = MVN(\mu_0, \Sigma_0)$

Assume: $\mu_0 = \vec{0}$, $\Sigma_0 = \sigma_0^2 \mathrm{I}$

$$\theta_{MAP} = \arg\min \log P(\theta|D) = \arg\min \mathcal{NLL}(\theta) + \log P(\theta)$$

We get

$$\theta_{MAP} = \arg\min \frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta) + \frac{1}{\sigma_0^2}\theta^T\theta$$

Question

What does this expression remind you of?

Using zero-mean Gaussian prior

$P(\theta) = MVN(\mu_0, \Sigma_0)$

Assume: $\mu_0 = \vec{0}$, $\Sigma_0 = \sigma_0^2 \mathrm{I}$

$$\theta_{MAP} = \arg\min \log P(\theta|D) = \arg\min \mathcal{NLL}(\theta) + \log P(\theta)$$

We get

$$\theta_{MAP} = \arg\min \frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta) + \frac{1}{\sigma_0^2}\theta^T\theta$$

Question

What does this expression remind you of?

Answer: Ridge Regression

Using zero-mean Laplace

Assume: each $\theta_i \sim \mathit{Laplace}(0, \sigma_0)$

Using zero-mean Laplace

Assume: each $\theta_i \sim \mathit{Laplace}(0, \sigma_0)$

$$\theta_{MAP} = \arg\min \log P(\theta|D) = \arg\min \mathcal{NLL}(\theta) + \log P(\theta)$$

Using zero-mean Laplace

Assume: each $\theta_i \sim \mathit{Laplace}(0, \sigma_0)$

$$\theta_{MAP} = \arg\min \log P(\theta|D) = \arg\min \mathcal{NLL}(\theta) + \log P(\theta)$$

We get

$$\theta_{MAP} = \arg\min \frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta) + \lambda|\theta|$$

Using zero-mean Laplace

Assume: each $\theta_i \sim \mathit{Laplace}(0, \sigma_0)$

$$\theta_{MAP} = \arg\min \log P(\theta|D) = \arg\min \mathcal{NLL}(\theta) + \log P(\theta)$$

We get

$$\theta_{MAP} = \arg\min \frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta) + \lambda|\theta|$$

Question

What does this expression remind you of?

Using zero-mean Laplace

Assume: each $\theta_i \sim \mathit{Laplace}(0, \sigma_0)$

$$\theta_{MAP} = \arg\min \log P(\theta|D) = \arg\min \mathcal{NLL}(\theta) + \log P(\theta)$$

We get

$$\theta_{MAP} = \arg\min \frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta) + \lambda|\theta|$$

Question

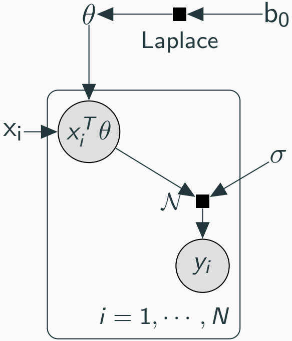
What does this expression remind you of?

Answer: LASSO Regression

Using Laplace prior

We can also use a Laplace prior on the weights, i.e.,

$$P(\theta) = \frac{1}{2b_0} \exp\left(-\frac{|x-\mu|}{b_0}\right)$$



Using Laplace prior

The MAP takes the form,

$$\theta_{MAP} = \arg\min \frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta) + \frac{1}{b_0}|\theta_i|$$

Using Laplace prior

The MAP takes the form,

$$\theta_{MAP} = \arg\min \frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta) + \frac{1}{b_0}|\theta_i|$$

Question

What does this expression remind you of?

Using Laplace prior

The MAP takes the form,

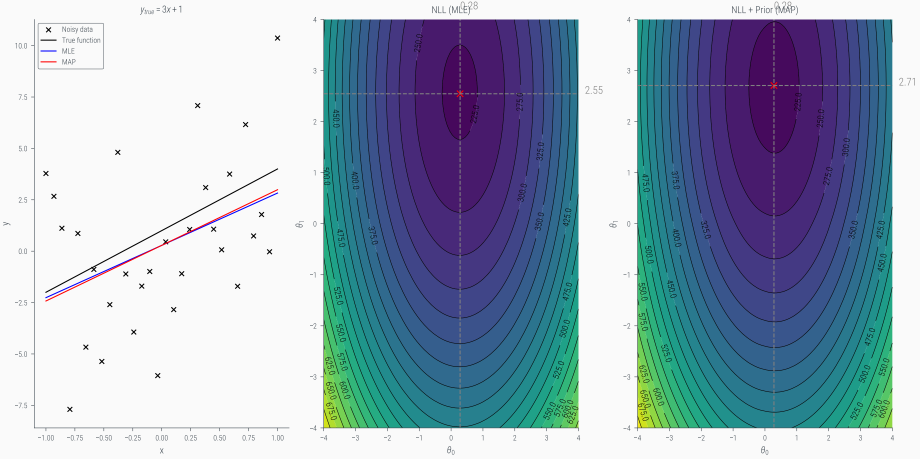
$$\theta_{MAP} = \arg\min \frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta) + \frac{1}{b_0}|\theta_i|$$

Question

What does this expression remind you of?

Answer: Lasso Regression

Using Laplace prior



Notebook

MAP for Logistic Regression

MLE for Logistic Regression

Consider a dataset $D = \{(x_1, y_1)...(x_N, y_N)\}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$ such that

$$P(y = 1|x) = \hat{y} = \frac{1}{1 + \exp(-X^T\theta)} = \sigma(X^T\theta)$$

Take $y \sim \text{Bernoulli}\left(\sigma(X^T\theta)\right)$

MLE for Logistic Regression

Consider a dataset $D = \{(x_1, y_1)...(x_N, y_N)\}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$ such that

$$P(y = 1|x) = \hat{y} = \frac{1}{1 + \exp(-X^T\theta)} = \sigma(X^T\theta)$$

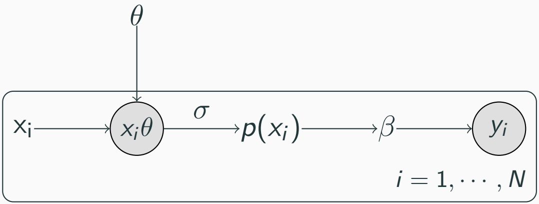
Take $y \sim \text{Bernoulli}\left(\sigma(X^T\theta)\right)$

The likelihood is given by

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \hat{y}_i{}^{y_i}(i - \hat{y}_i)^{1-y_i}$$

$$\implies \mathcal{L}\mathcal{L}(\theta) = \sum_{i=1}^{N} y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)$$

MLE for Logistic Regression



Binary Classification:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 X)}}$$

$$\therefore \mathcal{LL}(\theta) = \sum_{i=1}^{N} y_i \log p(x_i) + (1 - y_i) \log(1 - p(x_i))$$

Using zero-mean Gaussian prior

Considering a zero-mean Gaussian prior on the weights, i.e., $P(\theta) = \mathcal{N}(\theta|0, \sigma_0^2)$, the MAP is given by,

$$\theta_{MAP} = \arg\min \log(1 + \exp(-\theta^T X)) + \frac{1}{\sigma_0^2}\theta^T\theta$$

Using Laplace prior

Considering a Laplace prior on the weights, i.e., $P(\theta) = \prod_D \text{Laplace}(\theta_i|0, b_0) \propto \prod_D \exp(-\frac{1}{b_0}|\theta_i|)$, the MAP is given by,

$$\theta_{MAP} = \arg\min \log(1 + \exp(-\theta^T X)) + \frac{1}{b_0}|\theta|$$

MAP for Logistic Regression

Self-Study: Modify the code for Linear Regression to implement MAP for Logistic Regression.